

U.S. DEPARTMENT OF COMMERCE
National Technical Information Service

PB-253 576

REGRESSION USING 'HOCKEY STICK' FUNCTIONS

HEALTH EFFECTS RESEARCH LABORATORY

JUNE 1976

170018

EPA-600/1-76-024
June 1976

PB 253 576

Environmental Health Effects Research Series

REGRESSION USING "HOCKEY STICK" FUNCTIONS



REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U. S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

Health Effects Research Laboratory
Office of Research and Development
U.S. Environmental Protection Agency
Research Triangle Park, North Carolina 27711

RESEARCH REPORTING SERIES

Research reports of the Office of Research and Development, U.S. Environmental Protection Agency, have been grouped into five series. These five broad categories were established to facilitate further development and application of environmental technology. Elimination of traditional grouping was consciously planned to foster technology transfer and a maximum interface in related fields. The five series are:

1. Environmental Health Effects Research
2. Environmental Protection Technology
3. Ecological Research
4. Environmental Monitoring
5. Socioeconomic Environmental Studies

This report has been assigned to the ENVIRONMENTAL HEALTH EFFECTS RESEARCH series. This series describes projects and studies relating to the tolerances of man for unhealthful substances or conditions. This work is generally assessed from a medical viewpoint, including physiological or psychological studies. In addition to toxicology and other medical specialties, study areas include biomedical instrumentation and health research techniques utilizing animals—but always with intended application to human health measures.

TECHNICAL REPORT DATA (Please read instructions on the reverse before completing)		
1. REPORT NO. EPA-600/1-76-024	2.	3. RECIPIENT'S ACCESSION NO. PB-253 576
4. TITLE AND SUBTITLE Regression Using "Hockey Stick" Functions	5. REPORT DATE June 1976	
	6. PERFORMING ORGANIZATION CODE	
7. AUTHOR(S) Victor Hasselblad, John P. Creason and William C. Nelson	8. PERFORMING ORGANIZATION REPORT NO.	
9. PERFORMING ORGANIZATION NAME AND ADDRESS Statistics and Data Management Office Health Effects Research Laboratory U.S. Environmental Protection Agency Research Triangle Park, N.C. 27711	10. PROGRAM ELEMENT NO. 1AA601	
	11. CONTRACT/GRANT NO.	
12. SPONSORING AGENCY NAME AND ADDRESS Health Effects Research Laboratory Office of Research and Development U.S. Environmental Protection Agency Research Triangle Park, N.C. 27711	13. TYPE OF REPORT AND PERIOD COVERED	
	14. SPONSORING AGENCY CODE EPA-ORD	
15. SUPPLEMENTARY NOTES		
16. ABSTRACT The establishment of criteria for air pollutants requires that a threshold level be established below which no adverse health effects are observed. Since standard dose response curves, such as the logit or probit, assume an effect at all levels, a segmented function was developed. This function has zero slope up to a point, and then increases monotonically from that point. Thus the name "hockey stick" function. The increasing portion need not be linear; any function that can be fitted by "least squares" techniques will work. A method for computing confidence intervals is also given. Since the curve can be used as a dose response curve, some comparisons are made with the more conventional probit and logit curves. In general, the fit of the "hockey stick" curve is as good as either the logit or probit curve, even when the data originate from a logit or probit distribution.		
17. KEY WORDS AND DOCUMENT ANALYSIS		
a. DESCRIPTORS Biomathematics Biomedical measurements Mathematical logic Functions (mathematics) Special Functions	b. IDENTIFIERS/OPEN ENDED TERMS Hockey stick	c. COSATI Field/Group 12A 06T
18. DISTRIBUTION STATEMENT RELEASE TO PUBLIC	19. SECURITY CLASS (this Report) UNCLASSIFIED 20. SECURITY CLASS (if page) UNCLASSIFIED	21. NO. OF PAGES 16 22. PRICE 3.50

REGRESSION USING "HOCKEY STOCK" FUNCTIONS

By

Victor Hasselblad
John P. Creason
William C. Nelson
Statistics and Data Management Office
Health Effects Research Laboratory
U.S. Environmental Protection Agency
Research Triangle Park, N.C. 27711

U.S. ENVIRONMENTAL PROTECTION AGENCY
OFFICE OF RESEARCH AND DEVELOPMENT
HEALTH EFFECTS RESEARCH LABORATORY
RESEARCH TRIANGLE PARK, N.C. 27711

DISCLAIMER

This report has been reviewed by the Health Effects Research Laboratory, U.S. Environmental Protection Agency, and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

1. Introduction

In dose-response type studies, the logit or probit functions are usually applied to observed data to estimate the appropriate parameters of the response curve. Only the probit and logit functions have been seriously considered, as no theoretical basis has been suggested for other alternative functions. These two functions are practically indistinguishable except for very small or very large response regions. The rate of increase in response per unit increase in dose is frequently very small in these regions, and considerable difficulties are encountered in the determination of the function endpoints.

A major problem associated with the establishment of criteria for air pollutants is that of determining acceptable levels (thresholds) below which no effects are discernable. There is considerable support for the hypothesis that these thresholds do exist for chemical toxins ^(1,2,3), the demonstration of adaptive mechanisms being the most important. In this view, effects due to pollutants will appear only when changes induced in a person are beyond the capacity of that person's homeostatic mechanisms to overcome them. There will therefore be a threshold dose level, x_0 , below which there is no response other than the existing background response, y_0 . The dose-response relationship may then be described by regression functions of the form

$$y = y_0, x \leq x_0$$

$$y = y_0 + f(\underline{\theta}, x - x_0), x \geq x_0, f(\underline{\theta}, 0) = 0, \quad (1)$$

where y_0 , x_0 , and $\underline{\theta}$ are to be estimated from the available data. As mentioned above, considerable difficulties are encountered in trying to estimate x_0 using the logit or probit functions for $f(\underline{\theta}, x - x_0)$, so that alternative functions were considered. The particular case of equation (1) examined in this paper is what might be called a "hockey stick" function:

$$y = y_0, x \leq x_0$$

$$y = y_0 + b(x - x_0), x \geq x_0 \quad (2)$$

This function is a special case of two linear functions, with different slopes, that has been considered by Quandt⁽⁴⁾. His more general problem, which allowed for different variances as well, was solved by calculating the likelihood function for varying x_0 . The resulting tests for significance are based on the likelihood ratio criterion. Quandt's linear functions were not restricted to being continuous at the join point.

The more specific problem of obtaining least squares estimates of a segmented function which is continuous at the join points has been considered by Hudson⁽⁵⁾. The case of one join point is considered specifically, and Hudson's general method can easily be applied to the specific case of the "hockey stick" function.

2. Solutions of the Least Squares Equations

Suppose we have n pairs of observations (x_i, y_i) , and without loss of generality, assume that the x_i 's are ordered. Then the residual sum of squares, S , is

$$S(x_0) = \sum_{x_i < x_0} (y_i - y_0)^2 + \sum_{x_i \geq x_0}^n [y_i - y_0 - b(x_i - x_0)]^2 \quad (3)$$

The problem can be solved easily for fixed x_0 , giving a familiar looking set of normal equations:

$$\begin{pmatrix} n & \sum_{x_i > x_0}^n (x_i - x_0) \\ \sum_{x_i > x_0}^n (x_i - x_0) & \sum_{x_i > x_0}^n (x_i - x_0)^2 \end{pmatrix} \begin{pmatrix} y_0 \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{x_i > x_0}^n (x_i - x_0) y_i \end{pmatrix} \quad (4)$$

This could be done for each x_0 while x_0 is stepped in small increments from x_i to x_n . The value of x_0 , y_0 and b would be those values (not necessarily unique) giving the smallest sum of squares.

The method of Hudson, on the other hand, gives the exact minimum(s) in a finite number of steps. x_0 must either lie in an interval between x_j and x_{j+1} , or at one of the x_j 's, and so equations need be derived for these two cases only. If x_0 lies in the interval (x_j, x_{j+1}) then the values of x_0 , y_0 , and b are given by

$$y_0 = \frac{1}{j} \sum_{i=1}^j y_i \quad (5)$$

$$x_0 = \left[y_0 \sum_{i=j+1}^n x_i^2 + \left(\sum_{i=j+1}^n y_i \right) \left(\sum_{i=j+1}^n x_i^2 \right) - \left(\sum_{i=j+1}^n x_i \right) \left(\sum_{i=j+1}^n x_i y_i \right) - m y_0 \sum_{i=j+1}^n x_i^2 \right] / \left[\left(\sum_{i=j+1}^n x_i \right) \left(\sum_{i=j+1}^n y_i \right) - m \sum_{i=j+1}^n x_i y_i \right] \quad (6)$$

$$b = \left(\sum_{i=j+1}^n y_i - m y_0 \right) / \left(\sum_{i=j+1}^n x_i - m x_0 \right) \quad (7)$$

where $m = n - j$.

If x_0 lies at x_j , the values of y_0 and b are given by

$$y_0 \left[\left(\sum_{i=j+1}^n y_i \right) \left(\sum_{i=j+1}^n x_i^2 \right) - \left(\sum_{i=j+1}^n x_i \right) \left(\sum_{i=j+1}^n x_i y_i \right) \right] /$$

$$\left[n \sum_{i=j+1}^n x_i^2 - \left(\sum_{i=j+1}^n x_i \right)^2 \right] \quad (8)$$

$$b = \left(\sum_{i=j+1}^n y_i - n y_0 \right) / \sum_{i=j+1}^n x_i$$

The residual sum of squares can be computed for the $n-1$ intervals and n x_i 's. The absolute minimum sum of squares can always be determined in at most $2n-1$ steps.

It is clear that the problem is symmetric in that the function

$$y = y_0, \quad x \geq x_0$$

$$y = y_0 + b(x_0 - x), \quad x \leq x_0$$

can be fitted by making the transformation

$$z = -x.$$

3. Confidence Intervals for x_0

Hudson also gives methods for computing confidence intervals. This involves looking at "likelihood regions", which in this case are intervals such that

$$S(x'_0) \leq (1 + \delta) S(x_0)$$

where $S(x_0)$ is the residual sum of squares as a function of the join point. The value of δ can be approximately determined from F tables with 1 and $n-3$ degrees of freedom. As pointed out by Hudson and Quandt, this interval for

x_0 need not be connected, although the function $S(x_0)$ is continuous. It is also possible for x_1 and/or x_n to be contained in the interval. If x_1 is in the interval, then the "hockey stick" is not a significantly better fit than a straight line. If x_n is in the interval then no meaningful relationship between x and y has been discovered.

4. Examples

The first example consists of data artificially generated from the equation

$$y = 2 \text{ for } x \leq 4.5$$

$$y = x - 2.5 \text{ for } x \geq 4.5$$

The variance, σ_e^2 , is 0.25.

The points are

x	1	2	3	4	5	6	7	8	9	10
y	1.058	2.186	2.590	1.473	2.419	3.432	5.016	6.045	6.184	7.372

Figure 1 shows the points and the least squares fit, which is

$$y = 1.827 \text{ for } x \leq 4.158$$

$$y = .973 x - 2.218 \text{ for } x \geq 4.158$$

The confidence interval ($\alpha = .05$) for x_0 is (3.202, 5.155). $\hat{\sigma}_e^2 = .2956$.

A graph of the sum of squares, S , as a function of the join point, x_0 , is given in Figure 2.

The second example is taken from a 3 year study of 112 student nurses in Los Angeles ⁽⁶⁾. Each day the nurses filled in a health questionnaire that included questions on eye discomfort, headache, fever, cough, etc. The dependent variable was the percent of nurses who experienced

eye discomfort without experiencing fever. The independent variable was the maximum hourly oxidant, expressed in pphm.

A "hockey stick" curve was fitted to the data. The least squares fit gave

$$\begin{aligned} y &= 5.77 && \text{for } x < 14.67 \\ y &= 5.77 + .617(x - 14.67) && \text{for } x \geq 14.67 \end{aligned} \quad (7)$$

The 95 percent confidence limits for x_0 ($=14.67$) were (13.25, 16.37). The mean square error was 13.59. The relatively narrow confidence interval resulted from the large number of points (867) and the consistency of the data.

5. Comparison with Probit and Logit Curves

Although the use of a "hockey stick" curve provides a convenient method for hypothesis testing, this does not mean that the curve is a good fit to dose-response data. For this reason the "hockey stick" was compared with the probit and logit curves for data which was simulated from probit and logit curves. The results of these simulations are given in Tables 1 and 2. The measure of goodness of fit is the standard R^2 used in regression analysis.

There are several factors which make the comparison of the various curves difficult. The data was simulated by starting with a known probit (or logit) curve. 10, 20, 50 or 100 equally spaced points were selected from the 0th to 50th percentile points. At each point, sample sizes of

10, 20, 50 or 100 were generated using pseudo-random numbers. The parameters of the probit (or logit) curve and the hockey stick were estimated. Thus the generation of the data favors the probit (or logit) curve.

On the other hand, the criterion of fit is the total sum of squares (or standard R^2), which favors the hockey stick, since the probit (or logit) curves can be thought of weighted least squares fits. In addition, the hockey stick is a three parameter curve, whereas the probit and logit curves both have only two parameters. In spite of these difficulties, Tables 1 and 2 give a crude comparison of the fits of the hockey stick with the probit and logit curves.

The simulations were done for dose values ranging up to the 50th percentile (LD_{50}) of the distribution, since air pollution health data rarely goes beyond the 50th percentile. Under these restrictions, there was little or no difference between the fit of the "hockey stick" vs. either the probit or logit curve. This was true for 10 to 100 doses (K) with 10 to 100 subjects (N) per dose. All curves had better R^2 for increasing K and N, which is to be expected. Additional simulations were made using a maximum dose at the 25th percentile and a maximum dose at the 75th percentile. These simulations all showed a similar pattern.

6. Discussion

Although the fitting of a "Hockey stick" function to data is not a particularly difficult problem, there are at least two items which should be noted. First, this simple departure from linear regression gives a sum of squares function that may not have a first derivative with respect to x_0 at several points, a fact amply demonstrated in Figure 2 from our earlier example. This alone is enough to create problems for many non-linear least-squares fitting computer programs. Secondly, the implication of the comparison to the probit and logit curves is of interest. Currently there is a great deal of discussion about the extrapolation of dose-response curves to very low doses. If 100 observations at each of 100 points cannot distinguish between a "hockey stick" and a probit or logit curve, it is clear that the resolution of this problem strictly through large scale sampling is not feasible.

In summary, the "hockey stick" curve provides a convenient method for estimation and hypothesis testing in low-dose and/or high dose regions of dose response curves, the estimation procedures are simple and straight-forward, and it's fit to the data appears to be indistinguishable from that of the standard logit or probit curves. The use of the "hockey stick" function is definitely not a major breakthrough in curve fitting. It does, however, offer a means of testing for a threshold level that is not available using standard dose-response curves.

Table 1. Comparison of Fit for Hockey
Stick vs. Probit for Data from
K Binomials, $B(p, N)$, Where P Follows a Probit Curve
From 0.0 to 0.5. Data Based on 100 Simulations.

K	N	Variance About Hockey Stick	Variance About Probit Curve	Total Variance	R ² -Hockey	R ² -Probit
10	10	.10782	.11542	.38341	.719	.699
10	20	.05783	.05144	.29695	.805	.827
10	50	.02500	.02161	.26661	.906	.919
10	100	.01224	.01177	.25115	.951	.953
20	10	.26031	.24980	.70066	.628	.643
20	20	.14652	.12803	.56094	.739	.772
20	50	.05432	.05137	.49981	.891	.897
20	100	.02824	.02524	.46530	.939	.946
50	10	.65765	.67091	1.72323	.618	.611
50	20	.35227	.35072	1.39590	.748	.749
50	50	.13792	.12940	1.18444	.884	.891
50	100	.07339	.06290	1.11057	.934	.943
100	10	1.35167	1.34766	3.41015	.604	.605
100	20	.69290	.67612	2.74174	.747	.753
100	50	.28902	.26974	2.34373	.877	.885
100	100	.15862	.13646	2.18693	.927	.938

Table 2. Comparison of Fit for Hockey Stick vs. Logit for Data from K Binomials, $B(p, N)$, where P Follows a Logit Curve From 0.0 to 0.5. Data Based on 100 Simulations.

K	N	Variance About Hockey Stick	Variance About Logit Curve	Total Variance	R ² -Hockey	R ² -Logit
10	10	.13660	.16436	.31987	.573	.486
10	20	.06719	.07930	.22726	.704	.651
10	50	.02726	.03157	.17569	.845	.820
10	100	.01468	.01652	.15492	.905	.893
20	10	.20790	.33810	.57319	.463	.410
20	20	.16589	.18018	.44228	.625	.593
20	50	.06876	.07357	.34107	.798	.784
20	100	.03338	.03429	.29589	.887	.884
50	10	.85596	.89600	1.46710	.417	.389
50	20	.46038	.47988	1.10606	.584	.566
50	50	.17860	.18178	.81819	.782	.778
50	100	.09058	.09061	.72769	.876	.875
100	10	1.74630	1.80502	2.93453	.405	.385
100	20	.92286	.94463	2.18506	.578	.568
100	50	.37891	.38104	1.65716	.753	.770
100	100	.19615	.19270	1.44919	.865	.867

TECHNICAL REPORT DATA (Please read Instructions on the reverse before completing)		
1. REPORT NO. EPA-600/1-76-024	2.	3. RECIPIENT'S ACCESSION NO.
4. TITLE AND SUBTITLE Regression Using "Hockey Stick" Functions	5. REPORT DATE June 1976	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Victor Hasselblad, John P. Creason and William C. Nelson	8. PERFORMING ORGANIZATION REPORT NO.	
9. PERFORMING ORGANIZATION NAME AND ADDRESS Statistics and Data Management Office Health Effects Research Laboratory U.S. Environmental Protection Agency Research Triangle Park, N.C. 27711	10. PROGRAM ELEMENT NO. 1AA601	11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS Health Effects Research Laboratory Office of Research and Development U.S. Environmental Protection Agency Research Triangle Park, N.C. 27711	13. TYPE OF REPORT AND PERIOD COVERED	
	14. SPONSORING AGENCY CODE EPA-ORD	
15. SUPPLEMENTARY NOTES		
16. ABSTRACT The establishment of criteria for air pollutants requires that a threshold level be established below which no adverse health effects are observed. Since standard dose response curves, such as the logit or probit, assume an effect at all levels, a segmented function was developed. This function has zero slope up to a point, and then increases monotonically from that point. Thus the name "hockey stick" function. The increasing portion need not be linear; any function that can be fitted by "least squares" techniques will work. A method for computing confidence intervals is also given. Since the curve can be used as a dose response curve, some comparisons are made with the more conventional probit and logit curves. In general, the fit of the "hockey stick" curve is as good as either the logit or probit curve, even when the data originate from a logit or probit distribution.		
17. KEY WORDS AND DOCUMENT ANALYSIS		
a. DESCRIPTORS Biomathematics Biomedical measurements Mathematical logic Functions (mathematics) Special Functions	b. IDENTIFIERS/OPEN ENDED TERMS Hockey stick	c. COSATI Field/Group 12A 06T
18. DISTRIBUTION STATEMENT RELEASE TO PUBLIC	19. SECURITY CLASS. (This Report) UNCLASSIFIED 20. SECURITY CLASS. (If is pages) UNCLASSIFIED	21. NO. OF PAGES 16 22. PRICE 3.50